Similar Classifiers and VC Error Bounds

Eric Bax[*]

June 6, 1997

Abstract

We improve error bounds based on VC analysis for classes with sets of similar classifiers. We apply the new error bounds to separating planes and artificial neural networks.

Key words machine learning, learning theory, generalization, Vapnik-Chervonenkis, separating planes, neural networks.

[*] Computer Science Department, California Institute of Technology 256-80, Pasadena, California, 91125 (eric@cs.caltech.edu).

1 Introduction

Our machine learning framework has the following structure. There is an unknown boolean-valued target function, and there is a distribution over the input space of the function. For example, the input space could consist of images from a weather satellite, and the output could be 1 if the image contains a hurricane and 0 otherwise.

We have a set of in-sample data with inputs drawn according to the input distribution and outputs determined by the target function. The test example inputs will also be drawn according to the input distribution. We know the number of test example inputs, but not the inputs themselves. Our goal is to find a classifier function with a low error rate on the test inputs. (The error rate is the fraction of examples for which the classifier and the target function disagree.)

We select a class of classifiers without reference to the data. We use the in-sample data to select a classifier from the class, e.g., through training or data-fitting by other means. Vapnik-Chervonenkis analysis [17, 1, 3] provides probabilistic bounds on the test error of the chosen classifier. In this paper, we improve these bounds for several classes.

In the next section, we derive VC-type bounds. First, we derive a bound for a single classifier chosen without reference to the data. Then we derive a uniform bound for several classifiers. Since we choose our classifier with reference to the data, the single-classifier bound does not apply. However, the uniform bound over the class applies to the trained classifier because the class is chosen without reference to the data, and the trained classifier is in the class.

In the following section, we improve the uniform bound for classes with similar classifiers. First, we derive nearly uniform bounds over the class, i.e., bounds that allow a limited number of single-classifier bound failures. Then, we show that for pairs of similar classifiers, a bound for one classifier implies a slightly looser bound for the other. If every classifier has enough similar classifiers, then the nearly uniform bounds imply slightly looser uniform bounds because the classifiers for which the "direct" bound fails are "covered" by at least one of their neighbors for which the "direct" bound succeeds.

Next, we prove that the new bound applies to the class of separating planes. We show that the class has sufficiently similar classifiers to improve error bounds. Also, we extend the result for separating planes to many classes of artificial neural networks.

2 Uniform Error Bound

We derive a uniform error bound by generalizing the bound at the heart of the original VC paper [17]. Let d be the number of in-sample examples, and let d' be the number of test examples. Let ν represent in-sample error, and let ν' represent test error.

We begin with a test error bound for a single classifier chosen without reference to the data. For the moment, condition the bound on a given multiset of $n = d + d'$ inputs composing the in-sample and test inputs. Since the inputs are drawn i.i.d., each partition of the inputs into in-sample and test data is equally likely. Let w be the number of inputs for which the classifier produces the incorrect output. The probability that the in-sample error is $\frac{c}{d}$ is

$$\binom{b}{d}^{-1}\binom{w}{c}\binom{b-w}{d-c} \tag{1}$$

If the validation error is $\frac{c}{d}$, then the test error is $\frac{w-c}{d'}$. So

$$\Pr\{\nu'_m \geq \nu_m + \epsilon | w\} = \sum_{\{c|\frac{w-c}{d'} \geq \frac{c}{d} + \epsilon\}} \binom{b}{d}^{-1}\binom{w}{c}\binom{b-w}{d-c} \tag{2}$$

Bound by maximizing over w.

$$\Pr\{\nu'_m \geq \nu_m + \epsilon\} \leq \max_{w \in \{0,\ldots,b\}} \Pr\{\nu'_m \geq \nu_m + \epsilon | w\} \tag{3}$$

Refer to the bound as $B(\epsilon)$.

The probabilities are over partitions of the inputs into in-sample and test examples, so the bound does not apply if the classifier is chosen with reference to the partition represented by our in-sample and test data. This information is contained in our in-sample data, so the bound is not valid for a trained classifier.

Next, we derive a uniform bound over a set of classifiers chosen without reference to the partition. Since the bound is uniform, it applies to the individual classifier chosen by training. Hence, we use a uniform bound as a bound for the single trained classifier.

For a given multiset of $n = d + d'$ inputs, let M be the number of classifiers in the set. Let ν_m and ν'_m represent the in-sample and test errors of classifier m.

Let U be the set of partitions. Let F_m be the set of partitions for which the single-classifier bound fails for classifier m, i.e., for which $\nu'_m \geq \nu_m + \epsilon$. Let $|F_m|$ be the probability mass in F_m, i.e., the probability of single-classifier bound failure for classifier m. By (3),

$$|F_m| \leq B(\epsilon) \tag{4}$$

The probability that at least one single-classifier bound fails, $|F_1 \cup \ldots \cup F_M|$, is maximized when $F_1, \ldots, F_M$ are disjoint. In this case,

$$\Pr\{\nu'_1 \geq \nu_1 + \epsilon \text{ or } \ldots \text{ or } \nu'_M \geq \nu_M + \epsilon\} = |F_1 \cup \ldots \cup F_M| = |F_1| + \ldots + |F_M| \leq MB(\epsilon) \tag{5}$$

Hence, for a trained classifier selected from the set,

$$\Pr\{\nu' < \nu + \epsilon\} \geq 1 - MB(\epsilon) \tag{6}$$

This is the VC-type test error bound.

3 Improved Uniform Error Bound

3.1 Nearly Uniform Bounds

What is the probability that at least two single-classifier bounds fail? The probability is maximized when each partition in a failure set is in exactly two failure sets. In this case

$$|F_1 \cup \ldots \cup F_M| = \frac{1}{2}(|F_1| + \ldots + |F_M|) \leq \frac{M}{2}B(\epsilon) \tag{7}$$

The probability of k or more failures is maximized when each partition in the failure set is in exactly k failure sets. In this case

$$|F_1 \cup \ldots \cup F_M| = \frac{1}{k}(|F_1| + \ldots + |F_M|) \leq \frac{M}{k}B(\epsilon) \tag{8}$$

3.2 Rates of Disagreement and Error Bounds

Suppose a pair of classifiers, g and $g_{\&}$, have no more than r disagreements over the $n = d + d'$ inputs. For a partition, let γ be the rate of disagreement over the in-sample inputs, and let δ be the rate of disagreement over the test inputs. Suppose the single-classifier bound holds for $g_{\&}$, i.e., $\nu'_{\&} < \nu_{\&} + \epsilon$. Then

$$\nu'_{\&} - \nu' + \nu' < \nu_{\&} - \nu + \nu + \epsilon \tag{9}$$

which implies

$$\nu' < \nu + (\nu_{\&} - \nu) + (\nu' - \nu'_{\&}) + \epsilon \tag{10}$$

The difference in error rates is no greater than the rate of disagreement. So

$$\nu' < \nu + \gamma + \delta + \epsilon \tag{11}$$

Let $D = \min(d, d')$. Assuming $r \leq D$, $\gamma + \delta \leq \frac{r}{D}$. (The maximizing case has all disagreements in the smaller set of inputs.) This implies

$$\nu' < \nu + \frac{r}{D} + \epsilon \text{ for } r \leq D \tag{12}$$

Hence,

$$\nu'_{\&} < \nu_{\&} + \epsilon \quad \Rightarrow \quad \nu' < \nu + \frac{r}{D} + \epsilon \tag{13}$$

For a given set of $n = d + d'$ inputs, suppose each classifier in $g_1, \ldots, g_M$ has at least $k - 1$ "neighbors" with r or fewer disagreements. From (8), the probability of k or more single-classifier bound failures is at most $\frac{M}{k}B(\epsilon)$. Thus, the probability that each classifier has at least one classifier $g_{\&}$ among itself and its neighbors with $\nu'_{\&} < \nu_{\&} + \epsilon$ is at least $1 - \frac{M}{k}B(\epsilon)$. So

$$\Pr\{\forall m \in \{1, \ldots, M\} | \nu'_m < \nu_m + \frac{r}{D} + \epsilon\} \geq 1 - \frac{M}{k}B(\epsilon) \tag{14}$$

This bound is uniform over the set of classifiers, so it applies to the classifier selected by training. Hence,

$$\Pr\{\nu' < \nu + \frac{r}{D} + \epsilon\} \geq 1 - \frac{M}{k}B(\epsilon) \tag{15}$$

Compare bounds (6) and (15). Note that (6) is a special (trivial) case of (15), with $k = 1$ and $r = 0$. Bound (15) is strong when k is large and r is small, i.e., when there are large neighborhoods of classifiers with small rates of disagreement.

3.3 The Bounds in the VC Framework

Now, we extend bound (14) to the full VC framework, with infinite sets of classifiers and no conditioning on the inputs. Let G be a class of classifiers. For a given multiset $\{\mathbf{x}\}$ of n inputs, let $M(\{\mathbf{x}\})$ be the number of distinct output patterns generated by classifiers in the infinite set. For each partition of the inputs into in-sample and test sets, the in-sample and test errors are the same for all classifiers with the same output pattern. For each output pattern, select a representative classifier. Uniform bounding over the representative classifiers is equivalent to uniform bounding over all classifiers in the infinite set. Hence,

$$\Pr\{\forall g \in G | \nu' < \nu + \frac{r}{D} + \epsilon | \{\mathbf{x}\}\} \geq 1 - \frac{M(\{\mathbf{x}\})}{k}B(\epsilon) \tag{16}$$

Next, we remove the conditioning of the bound on the multiset $\{\mathbf{x}\}$, deriving distribution-free bounds. Assume that the conditions on k and r hold for all $\{\mathbf{x}\}$. Let $p(\{\mathbf{x}\})$ be the p.d.f. of multiset $\{\mathbf{x}\}$, given by the underlying input distribution.

$$\Pr\{\forall g \in G | \nu' < \nu + \frac{r}{D} + \epsilon\} = \int_{\{\mathbf{x}\}} \Pr\{\forall g \in G | \nu' < \nu + \frac{r}{D} + \epsilon | \{\mathbf{x}\}\} p(\{\mathbf{x}\}) d\{\mathbf{x}\} \tag{17}$$

Use bound (16).

$$\geq \int_{\{\mathbf{x}\}} (1 - \frac{M(\{\mathbf{x}\})}{k}B(\epsilon)) p(\{\mathbf{x}\}) d\{\mathbf{x}\} \tag{18}$$

Let $m(n) = \max_{\{\mathbf{x}\}} M(\{\mathbf{x}\})$,i.e., $m(n)$ is the maximum number of distinct output patterns generated by classifiers in G. (This number is known as the growth function [17].)

$$\geq \int_{\{\mathbf{x}\}} (1 - \frac{m(n)}{k}B(\epsilon)) p(\{\mathbf{x}\}) d\{\mathbf{x}\} \tag{19}$$

The probability bound is constant with respect to $\{\mathbf{x}\}$.

$$= (1 - \frac{m(n)}{k}B(\epsilon)) \int_{\{\mathbf{x}\}} p(\{\mathbf{x}\}) d\{\mathbf{x}\} \tag{20}$$

The integral is unity. Hence,

$$\Pr\{\forall g \in G | \nu' < \nu + \frac{r}{D} + \epsilon\} \geq 1 - \frac{m(n)}{k} B(\epsilon) \tag{21}$$

We restate the result as a theorem. Then we state a generalization and a variation.

Theorem 1 *Suppose that for each $\{\mathbf{x}\}$, for each classifier g in G, there are at least $k-1$ classifiers with distinct output patterns that have r or fewer disagreements with g over $\{\mathbf{x}\}$. Then*

$$Pr\{\forall g \in G | \nu' < \nu + \frac{r}{D} + \epsilon\} \geq 1 - \frac{m(n)}{k} B(\epsilon) \tag{22}$$

The conditions are violated if there are some $\{\mathbf{x}\}$ for which there are few output patterns, but the patterns have large rates of disagreement. We can weaken the conditions to cover these cases. Note that the bounds are uniform over $m(n)$ classifiers that fulfill the neighbor condition. If we can create the proper neighborhoods by adding $m(n) - M(\{\mathbf{x}\})$ or fewer classifiers, then the result holds.

Theorem 2 *Suppose that for each $\{\mathbf{x}\}$ there is a multiset of $m(n)$ or fewer output patterns with the following properties. First, the set contains every output pattern produced by the classifiers in G. Second, for each pattern there are at least $k-1$ other patterns with r or fewer disagreements. Then*

$$Pr\{\forall g \in G | \nu' < \nu + \frac{r}{D} + \epsilon\} \geq 1 - \frac{m(n)}{k} B(\epsilon) \tag{23}$$

Alternatively, we could use "central" classifiers [4]. In this case, it is not necessary for every representative classifier to have many neighbors. Instead, we require that a few classifiers form an r-disagreement covering of G. We can uniformly bound the error over the classifiers that form the covering by (5), then use them as $g_{\&}$'s to bound the others by (13).

Theorem 3 *Suppose that for each $\{\mathbf{x}\}$ there is a set of S or fewer central classifiers, and each classifier in G has r or fewer disagreements with some central classifier. Then*

$$Pr\{\forall g \in G | \nu' < \nu + \frac{r}{D} + \epsilon\} \geq 1 - SB(\epsilon) \tag{24}$$

The theorems, as stated, are distribution-free. If $\{\mathbf{x}\}$ is specified *a priori*, i.e., the validation and test inputs are known, then the results hold if the conditions are met for the $\{\mathbf{x}\}$ at hand.

The first theorem expresses the basic result. The second theorem makes the result more robust, allowing for instances $\{\mathbf{x}\}$ in which the neighbor bound condition does not hold. We will encounter such instances in the next section, when we use the second theorem on the class of separating hyperplanes. The third theorem is quite general. However, the central classifiers in its conditions may be difficult to specify in practice.

4 Application to Separating Planes

For inputs $\mathbf{x} \in R^d$, the separating plane classifiers are defined as follows:

$$f_{\mathbf{w},t}(\mathbf{x}) = \begin{cases} 1 & \mathbf{w} \cdot \mathbf{x} - t \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

Separating planes output 1 for all inputs on one side of a hyperplane and output 0 for all inputs on the other side. The growth function $m(n)$ for separating planes in d dimensions is [6]

$$m(n) = \begin{cases} 2^n & n \leq d+1 \\ 2\sum_{i=0}^{d} \binom{n-1}{i} & n > d+1 \end{cases} \tag{26}$$

Separating planes are a basic classifier set. For more information on separating planes, see [6, 5]. Perceptrons [6, 11, 12, 13, 9, 10] are closely related to separating planes. Multilayer artificial neural networks often employ separating planes in their final layer. Thus, they can be viewed as first mapping the input space into some intermediate space, then applying a separating plane. This idea is the basis for support vector machines [16].

4.1 Using Theorem 2

To use Theorem 2, we must show that for each $\{\mathbf{x}\}$ there is a multiset of $m(n)$ or fewer output patterns that includes all output patterns produced by separating planes and fulfills a neighbor condition. We prove a general neighbor condition – each output pattern has $2k$ or more neighbors with k or fewer disagreements for $k < \min(D, \frac{n}{2})$. Thus, we show

$$\Pr\{\nu' < \nu + \frac{k}{D} + \epsilon \;\forall g \in G\} \geq 1 - \frac{m(n)}{2k+1} B(\epsilon) \tag{27}$$

We denote output patterns by the set of inputs S_0 for which the output is 0. We begin with a one-dimensional input space, then we generalize to higher dimensions. For simplicity, we first assume that there are no duplicated inputs in $\{\mathbf{x}\}$.

Order the inputs $\mathbf{x}^1, \ldots, \mathbf{x}^n$ such that $\mathbf{x}^1 < \ldots < \mathbf{x}^n$. The output patterns produced by separating planes are

$$\emptyset, \{\mathbf{x}^1\}, \{\mathbf{x}^1, \mathbf{x}^2\}, \ldots, \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}, \{\mathbf{x}^2, \ldots, \mathbf{x}^n\}, \ldots, \{\mathbf{x}^{n-1}, \mathbf{x}^n\}, \{\mathbf{x}^n\} \tag{28}$$

Imagine these patterns arranged in order around a circle. Each pattern has one disagreement with the patterns on either side. Also, each pattern has k disagreements with the patterns k positions away. The patterns within k positions form a neighborhood of $2k$ patterns with k or fewer disagreements.

If two or more inputs share a position, then the separating planes will not produce the patterns in which S_0 includes only a portion of the points that share a position. Order the inputs such that $\mathbf{x}^1 \leq \ldots \leq \mathbf{x}^n$. Augment the pattern set to be (28), and the conditions of Theorem 2 are fulfilled.

For the higher-dimensional case, we once again begin with the assumption that all inputs are distinct. We reduce the higher-dimensional case to the one-dimensional case as follows. For each output pattern that can be produced by a separating plane, we create a corresponding separating plane from which all inputs are at distinct distances. To find neighboring patterns, we translate the plane along its normal vector, changing the classification of one input at a time.

Given $\{\mathbf{x}\}$ with distinct inputs, let the pattern set in Theorem 2 be the patterns produced by separating planes. Given a pattern that can be produced by a separating plane, construct a corresponding plane with all inputs at distinct distances as follows.

1. Start with any plane that generates the output pattern.

2. If $\mathbf{w} \cdot \mathbf{x} - t = 0$ for any $\mathbf{x} \in \{\mathbf{x}\}$, assign $t = t - \epsilon$, where $\epsilon = \frac{1}{2} \min_{\mathbf{x} \in S_0} |\mathbf{w} \cdot \mathbf{x} - t|$. The plane now has $\mathbf{w} \cdot \mathbf{x} - t \neq 0 \; \forall \mathbf{x} \in \{\mathbf{x}\}$, and it produces the same output pattern.

3. Let
$$\epsilon_i = \min_{\mathbf{x} \in \{\mathbf{x}\}} |\frac{\mathbf{w} \cdot \mathbf{x} - t}{x_i}| \tag{29}$$
for $i \in \{1, \ldots, d\}$. Note that ϵ_i is the minimum change in w_i that makes $\mathbf{w} \cdot \mathbf{x} - t = 0$ for some $\mathbf{x} \in \{\mathbf{x}\}$. Since $\mathbf{w} \cdot \mathbf{x} - t \neq 0 \; \forall \mathbf{x} \in \{\mathbf{x}\}$, $\epsilon_i > 0$.

4. Let
$$\delta_i = \min_{\mathbf{x}, \mathbf{x}' \in \{\mathbf{x}\} \text{ with } \mathbf{w} \cdot \mathbf{x} - t \neq \mathbf{w} \cdot \mathbf{x}' - t} |\frac{(\mathbf{w} \cdot \mathbf{x}' - t) - (\mathbf{w} \cdot \mathbf{x} - t)}{x_i - x_i'}| \tag{30}$$
Note that δ_i is the minimum change in w_i that makes two inputs with previously distinct distances from the plane now have the same distance. Also, note that $\delta_i > 0$.

5. If no pair of inputs $\mathbf{x}$ and $\mathbf{x}'$ are the same distance from the plane, then stop. Otherwise, let i be some index for which $x_i \neq x_i'$. Assign $w_i = w_i + \gamma_i$, where $\gamma_i = \frac{1}{2} \min(\epsilon_i, \delta_i)$. Now the pair of inputs have distinct distances. Repeat steps (3), (4), and (5).

Order the inputs such that $\mathbf{w} \cdot \mathbf{x}^1 - t < \ldots < \mathbf{w} \cdot \mathbf{x}^n - t$. As the parameter t is varied, the resulting hyperplanes produce patterns (28). By the proof for the one-dimensional case, the neighborhood conditions are fulfilled.

If there are duplicated inputs, first use a largest distinct subset to construct the neighbor-producing hyperplane as before. Then, treat the duplicated inputs as if they have slighlty different entries in the first dimension. When the neighbor producing hyperplane sweeps through a duplicated input, add the inputs to S_0 or remove them one at a time to form different output patterns. Order the duplicates. If $w_1 > 0$ and t is increasing, add the inputs to S_0 in order. If t is decreasing, remove the inputs from S_0 in the reverse order. Reverse the ordering if $w_1 < 0$. If we order the inputs such that $\mathbf{w} \cdot \mathbf{x}^1 - t \leq \ldots \leq \mathbf{w} \cdot \mathbf{x}^n - t$, and we order the duplicates as outlined, then patterns (28) are produced.

For this scheme to work, it is necessary that $w_1 \neq 0$ in the neighbor-producing plane. So if $w_1 = 0$, then assign $w_1 = \gamma_1$, as in step (5) of the construction procedure. It is also necessary that the output patterns produced with duplicated inputs are still present in the new scheme. These output patterns are still present – they are produced when the plane is not in contact with a duplicated input. Finally, it is necessary to show that, under the scheme, no more than $m(n)$ distinct output patterns are produced by the set of separating planes. To show this, we will show that all patterns produced by the scheme are also produced by an arrangement of the inputs – an alternative $\{\mathbf{x}\}$.

Let $\mathbf{x}^d$ be the position of duplicate inputs $\mathbf{x}^{d1}, \ldots, \mathbf{x}^{dm}$ in $\{\mathbf{x}\}$. We will show that, for some $\epsilon > 0$, all patterns produced by the scheme are produced by separating planes on

$$\{\mathbf{x}\}' = \{\mathbf{x}\} - \{\mathbf{x}^{d1}, \ldots, \mathbf{x}^{dm}\} \cup \{\mathbf{x}^{d1} + \frac{\epsilon}{m}\mathbf{e}^1, \mathbf{x}^{d2} + \frac{2\epsilon}{m}\mathbf{e}^1, \ldots, \mathbf{x}^{dm} + \epsilon\mathbf{e}^1\} \quad (31)$$

where $\mathbf{e}^1$ is the vector with first entry 1 and other entries 0. In other words, we will show that there is some space extending from $\mathbf{x}^d$ in the first dimension in which duplicate points can be placed along a line segment, and separating planes will produce the patterns generated by the scheme.

First, consider scheme patterns in which the duplicate inputs have the same output. These patterns can be produced by a separating plane that does not contact any input in $\{\mathbf{x}\}$. (The plane can be produced using steps (1) and (2) of the construction procedure.) Since the plane does not contact $\mathbf{x}^d$, there is some space between the plane and $\mathbf{x}^d$ in the direction of the first dimension.

Next, consider scheme patterns in which duplicate inputs have different outputs. Recall that the duplicates enter and leave S_0 in a prescribed order. Let this order correspond to the order of the "duplicates" along the line segment in $\{\mathbf{x}\}'$. Recall that the scheme produces different outputs for duplicates only when the neighbor-producing plane is in contact with $\mathbf{x}^d$.

We prove the result by showing that, for some $\epsilon > 0$, any pattern produced on $\{\mathbf{x}\} - \{\mathbf{x}^{d1}, \ldots, \mathbf{x}^{dm}\}$ by a plane through $\mathbf{x}^d$ can also be produced by a plane through $\mathbf{x}^d + \epsilon\mathbf{e}^1$. Hence, the scheme patterns can be produced on $\{\mathbf{x}\}'$ by sweeping a plane through positions $\mathbf{x}^{d1} + \frac{\epsilon}{m}\mathbf{e}^1, \mathbf{x}^{d2} + \frac{2\epsilon}{m}\mathbf{e}^1, \ldots, \mathbf{x}^{dm} + \epsilon\mathbf{e}^1$ without disturbing the classification of other inputs.

The original plane has $\mathbf{w} \cdot \mathbf{x}^d - t = 0$. Without loss of generality, assume

$w_1 < 0$. Now we construct the new plane. Assign $\mathbf{w}' = \mathbf{w}$ and $t' = t - \delta$, where

$$\delta = \frac{1}{2} \min_{\mathbf{x} \in S_0} |\mathbf{w} \cdot \mathbf{x} - t| \tag{32}$$

as in step (2) of the construction procedure. Note that $\delta > 0$ and that the output pattern remains the same. Now we find the $\epsilon > 0$ for which the new plane contacts $\mathbf{x}^d + \epsilon \mathbf{e}^1$.

$$\mathbf{w}' \cdot (\mathbf{x}^d + \epsilon \mathbf{e}^1) - t' = 0 \tag{33}$$

Recall $\mathbf{w}' = \mathbf{w}$ and $t' = t - \delta$.

$$\mathbf{w} \cdot (\mathbf{x}^d + \epsilon \mathbf{e}^1) - (t - \delta) = 0 \tag{34}$$

Factor out the original plane.

$$(\mathbf{w} \cdot \mathbf{x}^d - t) + w_1 \epsilon + \delta = 0 \tag{35}$$

Recall $(\mathbf{w} \cdot \mathbf{x}^d - t) = 0$.

$$\epsilon = \frac{\delta}{-w_1} \tag{36}$$

Since we assumed $w_1 < 0$, $\epsilon > 0$.

4.2 Using Theorem 3

These results can be developed in terms of the central classifiers of Theorem 3 as well as in terms of the neighborhoods of Theorem 2. Let P be the set of patterns S_0 produced by the class on $\{\mathbf{x}\}$. Let P_c be the subset of patterns with $|S_0| \bmod 2k+1 = c$. Let P_m be a subset with minimum cardinality. The set of central classifiers consists of one classifier that generates each pattern in P_m. Since elements of S_0 can be added or removed one at a time, each pattern in P is within k elements of a central classifier pattern. Note that $|P_m| \leq \frac{m(n)}{2k+1}$. By Theorem 3,

$$\Pr\{\forall g \in G | \nu' < \nu + \frac{k}{D} + \epsilon\} \geq 1 - \frac{m(n)}{2k+1} B(\epsilon) \tag{37}$$

Suppose we are bounding the test error of a single trained classifier in relation to its training error. If the test inputs are known, then the classifiers can be identified by pattern subset. For example, if a classifier has $|S_0| = m \pmod{2k+1}$, then it has the same pattern as a central classifier. Since Theorem 3 is based on uniform bounds over central classifiers,

$$\Pr\{\nu' < \nu + \epsilon\} \geq 1 - \frac{m(n)}{2k+1} B(\epsilon) \tag{38}$$

Likewise, for classifiers with patterns such that $|S_0| = m + c \pmod{2k+1}$ or $|S_0| = m - c \pmod{2k+1}$,

$$\Pr\{\nu' < \nu + \frac{c}{D} + \epsilon\} \geq 1 - \frac{m(n)}{2k+1} B(\epsilon) \tag{39}$$

5 Analysis

Now we analyze the neighbor bound to find the disagreement rate k that bounds the test error with maximum certainty. To simplify the analysis, we use the Hoeffding bound [7] $2e^{-\frac{1}{2}\epsilon^2 D}$ in place of the partition-based bound $B(\epsilon)$. (The Hoeffding bound is smooth, and it is often used in VC analysis [15].) We find that the best disagreement rate goes to $\mathrm{O}(\frac{1}{\epsilon})$ as the number of examples increases.

Our bound is

$$\Pr\{\forall g \in G | \nu' < \nu + \frac{k}{D} + \epsilon\} \geq 1 - \frac{m(n)}{2k+1} B(\epsilon) \tag{40}$$

Replace $B(\epsilon)$ by $2e^{-\frac{1}{2}\epsilon^2 D}$ to use the Hoeffding bound.

$$\Pr\{\forall g \in G | \nu' < \nu + \frac{k}{D} + \epsilon\} \geq 1 - \frac{m(n)}{2k+1} 2e^{-\frac{1}{2}\epsilon^2 D} \text{ for } \epsilon > 0 \tag{41}$$

Replace ϵ by $\epsilon - \frac{k}{D}$.

$$\Pr\{\forall g \in G | \nu' < \nu + \epsilon\} \geq 1 - \frac{m(n)}{2k+1} 2e^{-\frac{1}{2}(\epsilon - \frac{k}{D})^2 D} \text{ for } \epsilon > \frac{k}{D} \tag{42}$$

To find the optimal disagreement rate k, differentiate the confidence with respect to k, set the expression equal to 0, and solve.

$$\frac{\partial}{\partial k}(1 - \frac{m(n)}{2k+1} 2e^{-\frac{1}{2}(\epsilon - \frac{k}{D})^2 D}) = 0 \tag{43}$$

Note that $m(n)$ is constant with respect to k.

$$\frac{\partial}{\partial k} \frac{-2}{2k+1} e^{-\frac{1}{2}(\epsilon - \frac{k}{D})^2 D} = 0 \tag{44}$$

Differentiate.

$$\frac{2}{2k+1} e^{-\frac{1}{2}(\epsilon - \frac{k}{D})^2 D} [\frac{2}{2k+1} - \epsilon + \frac{k}{D}] = 0 \tag{45}$$

The expression outside the brackets is positive for $k > 0$.

$$\frac{2}{2k+1} - \epsilon + \frac{k}{D} = 0 \tag{46}$$

In the solution, $\epsilon > \frac{k}{D}$, so the k suggested by this analysis can be used in bound (42). As $D \to \infty$, the solution approaches $k = \frac{1}{\epsilon} - \frac{1}{2}$.

To see that the solution maximizes confidence, note that the solution has $k > \frac{1}{\epsilon} - \frac{1}{2}$, since $\frac{k}{D}$ is positive. Evaluate the derivative at $k = \frac{1}{\epsilon} - \frac{1}{2}$.

$$\frac{2}{2k+1} e^{-\frac{1}{2}(\epsilon - \frac{k}{D})^2 D} [\frac{k}{D}] \tag{47}$$

This is positive, so confidence increases as we move toward the solution k. Figures 1 and 2 show the behavior of confidence as k varies. The partition-based bound was used for the figures. In both figures, the optimal k is less than the value suggested by the analysis based on the Hoeffding bound.

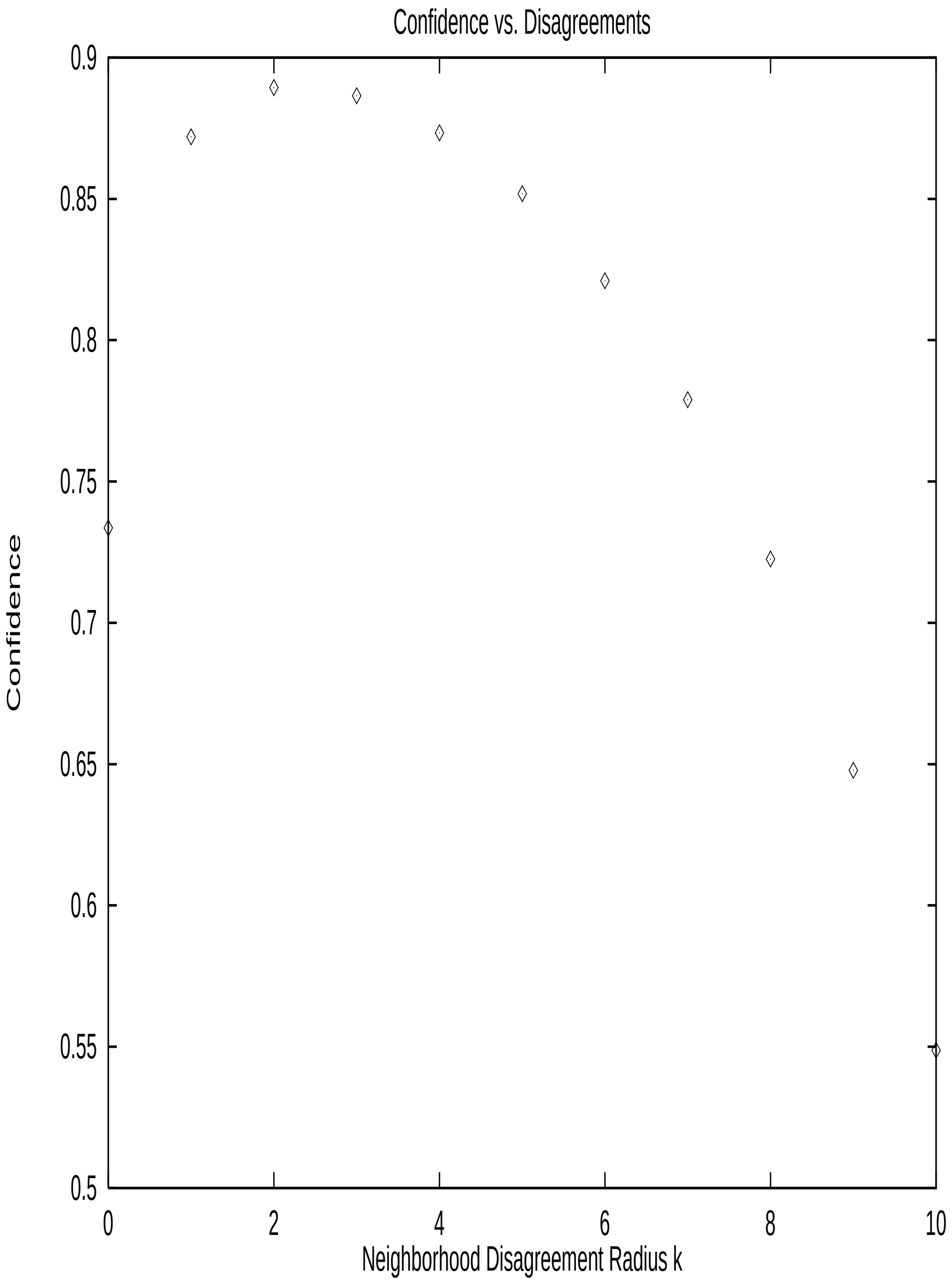

Figure 1: Confidence levels for test error bound (40) as the number of disagreements k among neighbors is varied. The bound is applied to separating planes in 5 dimensions, with $\epsilon = 0.18$, and 1000 examples in each of the training and test sets.

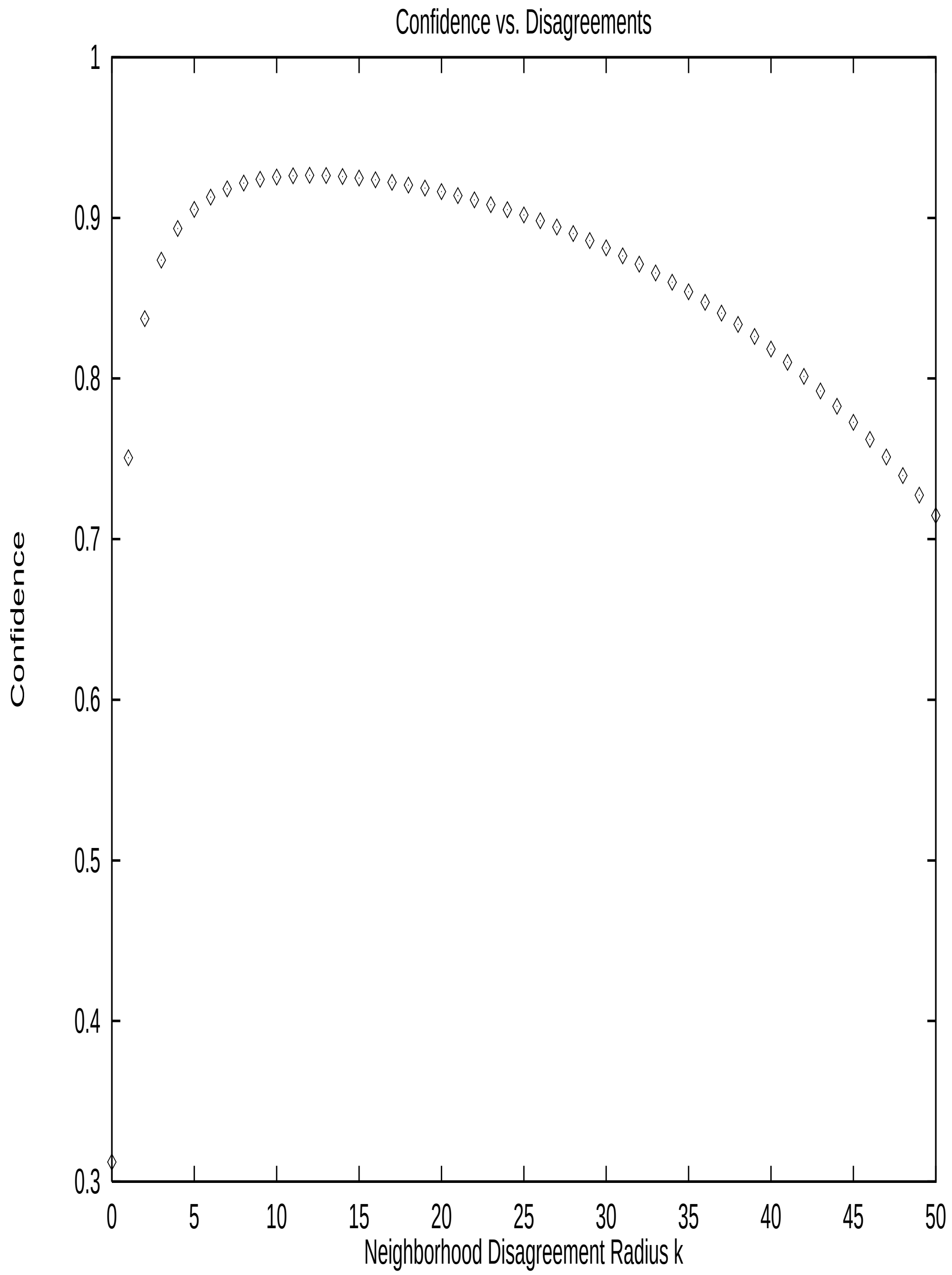

Figure 2: Confidence levels for test error bound (40) as the number of disagreements k among neighbors is varied. The bound is applied to separating planes in 1 dimension, with $\epsilon = 0.04$, and 5000 examples in each of the training and test sets.

6 Discussion

The bounds for separating planes also apply to multilayer neural networks in which the final layer acts as a separating plane over the space determined by the number of "hidden" units in the previous layer. To derive improved bounds for these networks, substitute the growth function of the particular architecture [2, 8] for $m(n)$ in bound (40).

It should not be difficult to extend the proof for separating planes to classes with other separating surfaces that can be continuously rotated and translated. Future challenges include developing neighbor bounds for other classes and improving the bounds for separating planes. (Cover [5] has results that may be useful for these purposes.)

Training by gradient descent requires classes with closely related classifiers. This work shows that the dense soup of classifiers necessary for learning by smooth descent on training error is not a disadvantage for generalization – classes containing sets of similar classifiers generalize better than their growth functions would indicate.

References

[1] Abu-Mostafa, Y.S. (1989). The Vapnik-Chervonenkis dimension: information versus complexity in learning. *Neural Computation*, 1 (3), 312-317.

[2] Baum, E. B., and Haussler, D. (1989). What size net gives valid generalization? *Neural Computation*, 1 (1), 151-160.

[3] Blumer, A., Ehrenfeucht, A., and Haussler, D. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36 (4), 929-965.

[4] Bax, E., Cataltepe, Z., and Sill, J. (1997). Alternative error bounds for the classifier chosen by early stopping. CalTech-CS-TR-97-08.

[5] Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14, 326-334.

[6] Duda, R., and Hart, P. (1973). Pattern Classification and Scene Analysis. John Wiley & Sons, Inc.

[7] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 13-30.

[8] Koiran, P. and Sontag, E. D. (1997). Neural networks with quadratic VC dimension. *Journal of Computer and System Sciences*, 54, 190-198.

[9] Nilsson, H. J. (1965). Learning Machines: Foundations of Trainable Pattern-Classifying Systems. New York: McGraw-Hill.

[10] Minsky, M., and Papert, S. (1969). Perceptrons: An Introduction to Computational Geometry. Cambrdige, MA: MIT Press.

[11] McCulloch, W. S., and Pitts, W. H. (1965). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5 (1943): 115-133.
reprinted in: McCulloch, W. S. (1965). *Embodiments of Mind*,(pp. 19-39). Cambrdige, MA: MIT Press.

[12] Rosenblatt, F. (1957). The perceptron – a perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory.

[13] Rosenblatt, F. (1962). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington, D.C.: Spartan Books.

[14] Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986). Parallel Distributed Processing. Cambrdige, MA: MIT Press.

[15] Vapnik, V. N. (1982). Estimation of Dependences Based on Empirical Data (p.31). New York: Springer-Verlag.

[16] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. New York: Springer-Verlag.

[17] Vapnik, V. N. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16, 264-280.